A Survey on Bioinformatics Tools

Bincy P Andrews[1], Binu A[2]
[1] Rajagiri School Of Engineering And Technology, Kochi, India
Email: bincymargret@gmail.com
[2] Rajagiri School Of Engineering And Technology, Kochi, India
Email: binu_a@rajagiritech.ac.in

***Abstract* — Bioinformatics may be defined as the field of science in which biology, computer science, and information technology merge to form a single discipline. Its ultimate goal is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned by means of bioinformatics tools for storing, retrieving, organizing and analyzing biological data. Also most of these tools possess very distinct features and capabilities making a direct comparison difficult to be done. In this paper we propose taxonomy for characterizing bioinformatics tools and briefly surveys major bioinformatics tools under each categories. Hopefully this study will stimulate other designers and experienced end users understand the details of particular tool categories/tools, enabling them to make the best choices for their particular research interests**.

***Index Terms* —Bioinformatics, tools, classifications.**

I. Introduction

Advances in technologies have been driven by demands in rapidly evolving scientific areas. Such a rapid growth occurred in the field of life sciences. This has in turn led to a high demand for computation tools that support the management, interpretation and analysis of the data generated by life science research. The field of Bioinformatics aims at addressing this demand. Complex applications are driving the need for new algorithms and tools to facilitate the access, analysis and interpretation of life science data. Bioinformatics enables the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. Hence it develops and improves methods for storing, retrieving, organizing and analyzing biological data. In this paper we present some of these tools by classifying them into several categories. Major kinds of classification includes sequence comparison tools, databases, data integration systems, Genome Browsers, Literature handling tools and programming languages. The focus of this survey is to provide a Comprehensive review of different bioinformatics tools by categorizing them effectively. Thus, the survey will help tool designers/developers and experienced end users understand the details of particular tool categories, enabling them to make the best choices for their particular research interests.

The entire paper is organized as follows; chapter II provides an overview of Sequence comparison tools. Chapter III provides details on Databases. Chapter IV provides insight of Genome browsers. In chapter V Literature handling tools are explained followed by chapter VI which describes programming languages for bioinformatics applications Followed by concluding remarks.

II. Sequence Comparison Tools

Sequence comparison tools[1][15] can be categorized into *sequence alignment,* including sequence homology searches and similarity scoring, *sequence filtering* methods, usually used for identifying, masking, or removing repetitive regions in sequences, and *sequence assembly* and *clustering methods*. *Blast* [2] is one of the most widely used bioinformatics tool developed in 1990. It can be accessed by any one from NCBI site and can be modified according to ones need. Its input data is in Fasta format and output data can be obtained in plain text, HTML or XML. Blast works on the principle of searching for localized similarities between the two sequences followed by short listing similar sequences for neighbourhood similarities. It searches for high number of similar local regions and gives the result after a threshold value is reached. This process differs from earlier softwares in which entire sequence was searched and compared which took a lot of time. It is used for many purposes like DNA mapping, comparing two identical genes in different species, creating phylogenetic tree. Blast is a threaded code and as such is limited to running with 8 Cpus or less, on one node. As the name suggests, *MPI-Blast* is an MPI code and is not limited to execution on a single node. Blast is more efficient than MPI-Blast in the sense that Blast expends fewer service-units for the task. The difference is considerable. Both Blast and MPI-Blast employ fragmented databases, one fragment for each thread for Blast and one fragment for each CPU for MPI-Blast. Blast remains efficient when deployed with more threads than Cpus but there is a limit to thread/CPU ratio. There is a hard limit to fragment size for both for Blast and MPI-Blast so when large databases are involved MPI-Blast is the only choice. *Fasta* [2] program was written in 1985 for comparing protein sequences only but it was later modified to conduct searches on DNA also. Fasta software uses the principle of finding the similarity between the two sequences statistically. It matches one sequence of DNA or protein with the other by local sequence alignment method. It searches for local region for similarity and not the best match between two sequences. Since it compares localized similarities at it can come up with a mismatches also. In a sequence Fasta takes a small part known as k-tuples where tuple can be from 1 to 6 and matches with k - tuples of other sequence and once a threshold value of matching is reached it comes up with the result. It is a program that is used to shortlist prospects of

DOI: 03.LSCS.2013.5. 48

matching sequence from a large number for full comparison as it is very fast. Difference between Blast and Fasta [8] is as follows:

- Blast is much faster than Fasta.
- Blast is much more accurate than Fasta.
- For closely matched sequences Blast is very accurate and for dissimilar sequence Fasta is better software.
- Blast can be modified according to the need but Fasta cannot be modified.
- Blast has to use Fasta input format to get the output data.
- Blast is much more versatile and widely used than Fasta.

HMMER [9] is used for searching sequence databases for homologs of protein sequences, and for making protein sequence alignments. It implements methods using probabilistic models called profile hidden Markov models. It creates an HMM model of a protein family. HMM contains both position-specific residue information and position specific gap penalties. It iteratively computes the gap penalties from the existing alignment, and then realigns the sequences according to the position specific gap penalties. Theoretically speaking it should be able to achieve its final alignment and model starting from unaligned sequences but in practice the process is more likely to settle on a good final model if the sequences are realigned. Most typically this is done by ClustalW. Compared to BLAST, FASTA, and other sequence alignment and database search tools based on older scoring methodology, HMMER aims to be significantly *more* accurate and *more* able to detect remote homologs because of the strength of its underlying mathematical models. In the past, this strength came at significant computational expense, but in the new HMMER3 project, HMMER is now essentially as fast as BLAST. Its advantages and disadvantages may be summarized as follows:

- Versus ClustalW as an alignment tool, HMMER provides a more objective placement of gaps. However, ClustalW provides the user with more capability to introduce secondary structural considerations.
- HMMER can be used as a search tool for additional homologues, whereas ClustalW cannot.
- HMMER is relatively dependent on sequences being divided up into domains in advance of model building.
- HMMER results are difficult to divorce from "expert" opinion, since both defining domains and defining the degree of divergence to be included in a protein family are based on preconceptions rather than computation.

Clustal W [16] [13] is a profile-based progressive alignment system. That means it first aligns the closest sequences to form a profile, representing the different residues at each position. It subsequently aligns sequence to profile and profile to profile using the information in the profile to help align distant families. Clustal W is the most commonly used algorithm after Psi-Blast, and the most commonly used algorithm to prepare seed alignments for HMMER. Clustal W is not a database searching system. It is used to align a set of sequences proposed to be homologues by another system. Gap penalties are more sophisticated with decreased penalties where sequences in alignment already have a gap, and an automatic capacity to avoid gaps in hydrophobic runs. However, they are still empirically adjusted, and the user may expect to do a lot of subjective fiddling if the alignment has very divergent sequences. Clustal W can accept a subset of the alignment prealigned by another system, and then carry on. It can also accept a secondary structure map, and adjust gap penalties to avoid gapping within secondary structure elements. It has got following advantages:

- Often implemented together with a Neighbor Joining phylogenetic analysis package, including Bootstrap analysis providing an easy interface.
- Implemented within Seaview multiple alignment editors.
- Default parameters widely accepted to produce an acceptable "objective" alignment on sequences greater than or equal to 25% identity.
- May do better than HMMER at relating divergent but well populated families.
- ClustalW will allow preserving prior alignments of two or more subsets as the global alignment is formed. The other methods will at most preserve the prior alignment of only one group.

Apart from advantages it also exhibits following disadvantages:

- Is easily confused by long stretches of unalignable sequences within otherwise well related sequences.
- Often produces a kind of false objectivity, where the user has fiddled with the program parameters to achieve a subjectively pleasing result, rather than just manually editing the alignment.
- Has no statistical evaluatory property. It will produce an alignment whether the provided sequences are related or not.
- Is less good than HMMER at adding a single divergent sequence to a family.

EMBOSS [11] [10] is "The European Molecular Biology Open Software Suite". It is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. It automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. It integrates a range of currently available packages and tools for sequence analysis into a seamless whole. It breaks the historical trend towards commercial software packages. It contains over 150 command-line tools for analyzing DNA/protein sequences that include pattern searching, phylogenetic analysis, data management, feature predictions, proteomics and more. Following are its features:

- Free
- Always available
- Command-line based
- Wide variety of programs

DOI: 03.LSCS.2013.5. 48

- Easy reformatting and parsing
- Remote database access

Glimmer[16] is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA. Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed. It has been used to annotate the complete genomes of over 100 bacterial species from TIGR and other labs. **PHYLIP** [16] (the PHYLogeny Inference Package) from the Dept. of Biology at the University of Washington is a package of programs for inferring phylogenies. It is available free over the Internet, and written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and executables are also distributed. It is most widely-distributed phylogeny package and third most frequently cited phylogeny package, after PAUP and MrBayes, and ahead of MEGA. PHYLIP is also the oldest widely-distributed package. It has been in distribution since October, 1980, and has over 28,000 registered users. It is regularly updated. **MrBayes** From School of Computational Science at the Florida State University is a program for the Bayesian estimation of phylogeny. Bayesian inference of phylogeny is based upon a quantity called the posterior probability distribution of trees. It is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. Since posterior probability distribution of trees is impossible to calculate analytically MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees. **T-Coffee** (**T**ree-based Consistency **O**bjective **F**unction **F**or alignment **E**valuation) [12] From Information Genomique et Structurale at Centre National de la Recherche Scientifique is a multiple sequence alignment software using a progressive approach. It generates a library of pairwise alignments to guide multiple sequence alignment. It can also combine multiple sequences alignments obtained previously. Its latest versions can use structural information from PDB files (3D-Coffee). It has advanced features to evaluate the quality of the alignments and some capacity for identifying occurrence of motifs (Mocca). It produces alignment in the aln format (Clustal) by default, but can also produce PIR, MSF, and FASTA format. The most common input formats are supported (FASTA, PIR). A notable feature of T-Coffee is its ability to combine different methods and different data types. In its latest version, T-Coffee can be used to combine protein sequences and structures, RNA sequences and structures. It can also run and combine the output of the most common sequence and structure alignment packages. **TIGR**[14][13] is used for assembling large shotgun DNA sequences. The TIGR Assembler overcomes several major obstacles to assembling such projects: the large number of pairwise comparisons required the presence of repeat regions, chimeras introduced in the cloning process, and sequencing errors. A fast initial comparison of fragments based on oligonucleotide content is used to eliminate the need for a more sensitive comparison between most fragment pairs, thus greatly reducing computer search time. Potential repeat regions are recognized by determining which fragments have more poten- tial overlaps than expected given a random distribution of fragments. Repeat regions are dealt with by increasing the match criteria stringency and by assembling these regions last so that maximum information from nonrepeat regions can be used. The algorithm also incorporates a number of constraints, such as clone length and the placement of sequences from the opposite ends of a clone. **GROMACS** (GROningen MAchine for Chemical Simulations) is a molecular dynamics package primarily designed for simulations of proteins, lipids and nucleic acids. It was originally developed in the Biophysical Chemistry department of University of Groningen, and is currently maintained by contributors in universities and research centers across the world. It is one of the fastest and most popular software packages available, and can run on CPUs as well as GPUs. It is free, open source released under the GNU General Public License. Following are its features:

- GROMACS is extremely fast due to algorithmic and processor-specific optimization, typically running 3-10 times faster than many simulation programs.
- GROMACS is operated via the command-line, and can use files for input and output.
- GROMACS provides calculation progress and ETA feedback, a trajectory viewer, and an extensive library for trajectory analysis.
- Support for different force fields makes GROMACS very flexible.
- It can be executed in parallel, using MPI or threads.
- GROMACS contains a script to convert molecular coordinates from a PDB file into the formats it uses internally.
- Once a configuration file for the simulation of several molecules has been created, the actual simulation run produces a trajectory file, describing the movements of the atoms over time which can be analyzed or visualized with a number of supplied tools.

In short functions of each of the tools mentioned above are provided in table 1. That is 12 tools followed by its function. It may be considered as a brief description of each of these tools for quick reference. Next chapter provides basic idea regarding databases which is the second category among bioinformatics tools.

III. Databases

Databases [1] can be categorized into protein sequence databases and DNA sequence databases. Table 2 provides an overview of databases in each of these categories.

A. Protein sequence databases

Protein sequences are the fundamental determinants of biological structure and function. The Protein database is a collection of sequences from several sources.

DOI: 03.LSCS.2013.5. 48

TABLE 1. BIOINFORMATICS TOOLS

TOOL	Function
HMMER	To identify homologous protein or nucleotide sequences founded by Sean Eddy.
NCBI BLAST	Is an algorithm for comparing primary biological sequence information
ClustalW	Is a widely used multiple sequence alignment computer program.
EMBOSS	is a free open source software analysis package
Glimmer	Is effective at finding genes in bacteria, archaea, and viruses.
Fasta	designed for protein sequence similarity searching
MrBayes	Is a program for Bayesian inference
Phylip	Is a free computational phylogenetics package of programs for inferring evolutionary trees
T_Coffee	To align sequences or to combine the output of your favorite alignment methods into one unique alignment
TIGR Assembler v2	TIGR is used for assembling large shotgun DNA sequences.
MPI-Blast	By efficiently utilizing distributed computational resources through database fragmentation, query segmentation, intelligent scheduling, and parallel I/O, mpiBLAST improves NCBI BLAST performance by several orders of magnitude
GROMACS	Is a molecular dynamics package primarily designed for simulations of proteins, lipids and nucleic acids.

The UniProt Consortium founded in 2007 organizes principal protein-sequence repositories in the public domain. Its databases consist of four components: Archive that provides a complete body of publically available protein sequence data without redundancy, Database of protein sequences with annotated sequences, Reference Clusters databases that provide non-redundant views of sequences at varying levels of resolution and Metagenomic and Environmental Sequences database for high-throughput genome analysis for unclassified organisms. **The National Center for Biotechnology Information (NCBI) GenBank** founded in 2000 is an international collection that includes data from the European Molecular Biology Laboratories and the DNA Databank of Japan. It includes translations of genome coding sections and cDNA sequences in GenBank sequences, UniProt, and the Protein Data Bank. Its DNA-centric view links protein and nucleotide sequences to various resources including the NCBI genome browser. It maintains species-specific sequence collections and provides BLAST formatted instances of the nr collection without duplication of sequences. The NCBI also provides Reference Sequence database that organizes protein sequence data in a genomic and taxonomical context. **The Protein Data Bank** is a principal data repository for protein and DNA three-dimensional structures established in 1970. The primary role of the PDB is to provide structural data. In the PDB format, records for atoms are contained within a hierarchical structure that supplies the atom names and coordinates are grouped by amino acid. PDB also provides information about the entry's chemistry, origin, and other details. **PROSITE** archives a large collection of biologically meaningful signatures expressed through regular-expression like patterns for short motif detection. Probabilistic profiles extend regular expressions in order to detect large domains that are more adaptive to variation in sequences.

B. DNA sequence databases

DNA sequence database is a collection of sequences from several sources. **GenBank** is a comprehensive database of publicly available DNA sequences for more than 300,000 named organisms maintained and distributed by NCBI. NCBI builds GenBank from several sources. Sources includes data from authors, the bulk submission of expressed sequence tag, genome survey sequence, whole genome shotgun and other high-throughput data from sequencing centers. The GenBank records are partitioned into divisions according to source organism or type of sequence. Records within the same division are packaged as a set of numbered files. GenBank release is offered in two formats flatfile format, and ASN.1 format. **EMBL** manages many databases containing biological information, including nucleic acid and protein sequences, and macromolecular structures. Data and bioinformatics services are made available freely to the scientific community. Among the databases provided, the most widely known are the EMBL Nucleotide Sequence Database. The EMBL database contains the results from sequencing experiments in laboratories together with the interpretation provided by the submitters, but with no or very limited review. The data in the EMBL database is gathered from large-scale genome sequencing projects, the European Patent Office, and direct submissions from individual scientists. Data in the EMBL Nucleotide Sequence Database are grouped into divisions, according to either the methodology used in their generation or taxonomic origin of the sequence source. Advantage of such a division becomes obvious when we wish to limit your database search or sequence similarity analysis to limited entries rather than the whole database.

IV. GENOME BROWSERS

Genome browsers [1] can be classified into three Web-based Genome Browsers, Standalone Annotation Browsers and Editors and Web-Based Synteny Browsers. **Web-based Genome Browsers** includes the University of California at Santa Cruz (UCSC) Genome Browser, Ensembl and NCBI Map Viewer [7]. Each uses a centralized model. It provides access to a large public database of genome data for many species and also integrates specialized tools. These public

DOI: 03.LSCS.2013.5.48

TABLE 2. DATABASES

DATABASE NAME	FOUNDED IN	PURPOSE
The UniProt Consortium	2007	organizes principal protein-sequence repositories in the public domain.
NCBI	2000	international collection that includes data from the European Molecular Biology Laboratories and the DNA Databank of Japan.
The Protein Data Bank	1970	data repository for protein and DNA three-dimensional structures
PROSITE	1988	archives a large collection of biologically meaningful signatures
GenBank	1982	Contains DNA sequences for more than 300,000 named organisms
EMBL	1981	Contains biological information, including nucleic acid and protein sequences, and macromolecular structures

browsers provide a valuable service to the research community by providing tools for free access. The Genome Browser at the UCSC provides access to genome assemblies obtained from the NCBI and external sequencing centers. Its main advantage is the speed and stability of its user interface. It is written in the C programming language and uses a combination of a MySQL database and highly optimized data serialization techniques. The Ensembl genome browser has feature-rich user interface. The sequence data are derived from EMBL and other external sources. It has got extensive annotation pipelines. The Ensembl browser supports extensive keyword searching. Since Ensembl user interface is very feature-rich it will not take some time to learn. The Map Viewer data is obtained from resources of the NCBI toolkit and also from the vast stores of sequence data and annotations in GenBank. Species coverage is higher than the other browsers; the map viewer currently has 106 species. It differs from others in terms of look and feel. Strength of the NCBI map viewer is its tight integration with other well-known NCBI resources. Like Ensembl and the UCSC browser, sequence similarity search capabilities are also available in the form of BLAST. **Standalone Annotation Browsers and Editors** offer many advantages to users. Some of these browsers include Apollo, IGB and Artemis**.** Apollo is a Java application**.** It has since been included in the GMOD project and several other organizations have adopted it for genome annotation. It supports a wide variety of formats for importing and exporting data and its main advantage as a browser is the ability for users to easily add and modify their own annotation data using a graphical tool and save changes to a remote database or a local file. Its disadvantage is the limit in total amount of sequence features that it can display at one time. Apollo will typically be looking at a small region in detail as opposed to large scale whole genome data. The Integrated Genome Browser is another Java based stand-alone browser. IGB will intelligently retrieve an appropriate amount of data of either a whole chromosome or only the viewed region. It works well with whole genome analysis results. IGB scrolls and zooms smoothly, even over large sequences with many features. IGB also has sophisticated analysis tools. Artemis is a Java based browser developed at the Sanger Institute. Artemis interface is simple and easy to master, but lacks the color and flashiness in some of the other standalone browsers. It allows the user to view and edit both sequence and annotation data. It is still under active development. **Web-Based Synteny Browsers** consists of GBrowse_syn and SynView .GBrowse_syn displays synteny data with a central reference genome. Data are stored for GBrowse_syn in separate databases. GBrowse-compatible database and alignment database. This arrangement gives the system the flexibility to change the primary sequence context for the display. The alignment database schema supports storage of information about insertions and deletions in the multiple sequence alignment. SynView uses the standard GBrowse distribution and a sophisticated GBrowse configuration file. It uses Perl callbacks to overlay a synteny view onto the GBrowse details panel. The synteny data are stored in standard GFF3Format. SynBrowse was written to take advantage of the BioPerl SynBrowse data are expressed as standard GFF2. Table 3 provides examples for each of these categories.

TABLE 3. BROWSERS

BROWSER	EXAMPLES
Web-based Genome Browsers	• UCSC • Ensembl • Map Viewer
Standalone Annotation Browsers and Editors	• Apollo • IGB • Artemis
Web-Based Synteny Browsers	• GBrowse_syn • SynView • SynBrowse

V. LITERATURE HANDLING TOOLS

It consists of literature mining tools [1] and literature databases. They are necessary to gather information on existing approaches of bioinformatics. Following sections describe each of them briefly.

A. Literature mining tools

Literature mining tools consists of Sentence and Entity Extraction Tools, Relation Extraction-Entity Linking Tools and Visualization Tools. **Sentence and Entity Extraction Tools** presents a ranked list of relevant sentences. Textpresso is an information extracting and processing package for biological

DOI: 03.LSCS.2013.5.48

literature associated with WarmBase, provides access to over 80,000 publications covering various model organisms. Collection of the full text of scientific articles is split into individual sentences. EbiMed allows searching a local copy of Medline. It searches retrieval results for sentences containing biomedical terminology and generates a table that displays proteins, GO annotations, drugs, and species extracted from sentences in retrieved abstracts. **Relation Extraction and Entity Linking Tools** identifies and extract associations between entities. The iHOP system organizes relevant sentences extracted from literature search results using gene/protein starting with a search for a gene or protein of interest a user will see search results in the form of sentences containing gene names. Search results provide an opportunity to navigate to information about all other genes and biomedical terms identified in the page and linked to external resources, and build an interaction network of genes/ proteins that co-occur in sentences. **Visualization Tools** has got visualization capabilities. Arizona Network Visualizer presents a framework for pathway-related knowledge integration and visualization. Semantic MEDLINE is another visual exploratory demonstration of the results of relations and entity extraction from 35 PubMed searches. Table 4 provides an overview of each of these categories.

TABLE 4. LITERATURE MINING TOOLS

TOOL	EXAMPLES
Sentence and Entity Extraction Tools	▪ Textpresso ▪ EbiMed
Relation Extraction and Entity Linking Tools	▪ iHOP ▪ Chilibot
Visualization Tools	▪ ARROWSMITH ▪ BITOLA

B. Literature databases

PubMed was developed by NCBI. It currently holds data for over 18 million articles from more than 34,000 journals dating back to the 1950s. Its resource is freely accessible and indexes the majority of life sciences articles. Records are enhanced with human-curated metadata, contain additional links to other resources, and are easily obtained via programmatic access and a web interface. PubMed archives and indexes the abstracts of articles and provides links from each article to a number of related resources. PubMed offers much functionality beyond the scope of its digital archive**. PubMed Central** was founded in 2000 is a digital archive which houses full text, open access literature in the life sciences. It is developed and managed by NCBI and benefits from strong integration with PubMed. **HighWire** Press, a division of the Stanford University Libraries, hosts nearly 5 million full text articles dating back to 1812 covering many science and social science topics. Programmatic access is not available but RSS feeds and email alerts are available. The Digital Repository Infrastructure Vision for European Research, DRIVER project was developed by an international consortium to manage and serve scientific information to European countries. Its aim is to provide a unified interface through which a user can search or browse content, while allowing the information to be remotely hosted by many institutions rather than archiving the information in a single institution. The **Web of Science** indexes a large number of articles – nearly one million, from about 8,700 journals – and has a powerful search interface. It is the resource of choice for statistics about the lives of articles after they are published. **arXiv** is a digital repository and pre-print server with over half a million open access articles in physics, mathematics, computer science, quantitative biology, and statistics, maintained by Cornell University and offers unrestricted access to its content. **CiteSeer** archives articles primarily in computer and information sciences and is a crucial resource in these fields. It archives full-text articles as Postscript and PDF files and is available without restriction. Table 5 summarizes each of these databases.

TABLE 5. LITERATURE DATABASES

TOOL	DESCRIPTION
PubMed	▪ Archives and indexes the abstracts of articles ▪ provides links from each article to a number of related resources.
PubMed Central	▪ It is a digital archive which houses full text, open access literature in the life sciences.
HighWire Press	▪ covers many science and social science topics
DRIVER Project	▪ Used to manage and serve scientific information to European countries
Web of Science	▪ covers the sciences and the arts
arXiv	▪ is a digital repository and pre-print server with over half a million open access articles in physics, mathematics, computer science, quantitative biology, and statistics
CiteSeer	▪ CiteSeer archives articles primarily in computer and information sciences and is a crucial resource in these fields

VI. PROGRAMING LANGUAGES

The **Perl** programming language gained popularity because of the ease of processing text files. It is an interpreted language, which allows for more rapid code, test and deploy processes. It also has got the rich set of available extensions, or modules, that provide pre-written functionality and has built a large web-accessible and searchable collection of freely available modules. The BioPerl project [5] is an effort to create a library of modules and routines in Perl that can be used to connect bioinformatics applications and datasets for rapid development of an application. Its aim is to abstract away the aspects of parsing data so that one can focus on accessing the information from analysis reports or raw data files. Typical use includes parsing FastA, GenBank, or EMBL format

DOI: 03.LSCS.2013.5. 48

sequences files, processing a BLAST report and generating GFF, General Feature Format files. **C & CPP** Provides open source libraries for building bioinformatics applications. They tend to be the fastest implementations for bioinformatics solutions since they are compiled to machine-executable code. The Ajax library is a C library that is part of the EMBOSS package and fills a variety of sequence and alignment needs. The libsequence library also provides a collection of C language routines for population genetics analyses. For the Cpp language, NCBI [3] has developed the general purpose NCBI Toolkit. There is a wide array of toolkit libraries and standalone applications for bioinformatics written in **Java**. Java is particularly good at providing graphical user interfaces. Due to optimization of the Java Virtual Machine in modern JDKs, applications offers much execution speed. It is a best choice for applications that needs high degree of reliability and robustness. The BioJava package provides rigorously designed object models for sequences, sequence features, alignments, dynamic programming algorithms, and for building classification algorithms such as support vector machines. The **Python** programming language has also gained much popularity in bioinformatics as a scriptable language. The BioPython project [6] provides a library of bioinformatics routines for sequence, structure, and alignment parsing in the Python programming language. The BioRuby project is also rapidly developing a bioinformatics library in the relatively new scripting and object-oriented programming language **Ruby**. The **R** software environment is a powerful platform for rapidly prototyping and assessing statistical analyses. R is a scripted language with a command-line interpreter that executes commands as they are typed, which together with the sophisticated plotting capabilities allows for quick successions of analysis refinement and result visualization. It can easily load algorithms at runtime that are coded in C for speed efficiency, and it has built-in support for creating packages without requiring much effort from the developer. R packages are organized in the CRAN repository. One of its projects includes BioConductor [4]. Table 6 shows various programming languages & their corresponding projects in the field of bioinformatics. Following table summarizes about each of these languages.

TABLE 6. PROGRAMMING LANGUAGES AND PROJECTS

LANGUAGE	PROJECT
Perl	BioPerl
Python	BioPython
R programming language	BioConductor
Ruby	BioRuby
Java	BioJava

VII. CONCLUSIONS

There has been a great deal of progress in development of new bioinformatics tools for infrastructure. In addition to development of new tools, user interfaces and documentation are available for familiarizing them in a user-friendly manner. Many of these tools and toolkits are open source, allowing them to be integrated into other applications and pipelines as reusable building blocks. It is necessary to understand about each and every tool available for its effective adoption by researchers. This paper provides details on most of the commonly used tools by categorizing them effectively. Future work will continue to expand this infrastructure and the types of software applications that can be used and provide more powerful and robust analysis tools for bioinformatics.

ACKNOWLEDGMENT

We are greatly indebted to the college management and the faculty members for providing necessary facilities and hardware along with timely guidance and suggestions for implementing this work.

REFERENCES

[1] David Edwards "Bioinformatics tools and application" Springer New York 2009
[2] Byoung-Tak Zhang and Chul Joo Kang "Bioinformatics Tools" http://bi.snu.ac.kr/Courses/g-ai01/bio-tools.pdf
[3] "Ncbi" http://www.ncbi.nlm.nih.gov/guide/sequence-ana lysis/
[4] "bioconductor" http://www.bioconductor.org
[5] "bioperl" http://www.bioperl.org
[6] "Biopython" https://www.biopython.org
[7] "Genegateway" http://www.ornl.gov/sci/techresources / Human_Genome /posters/chromosome/tools.shtml
[8] "Difference Between Blast and Fasta" http://www.differencebetween.com/differencebetween-blast-and-vsfasta/
[9] "Programming systems for protein family comparison" http://biochem.uthscsa.edu/~hs_lab/frames/molgen/tutor/compare.html
[10] "Bioinformatics Services."http://www.bioinformatics.wsu.edu/bioinfo_emboss.shtml
[11] "EMBOSS Manuals" http://manuals.bioinformatics.ucr. edu/home/emboss
[12] "T-Coffee" http://en.wikipedia.org/wiki/T-Coffee
[13] "JCVI: TIGR Assembler: A New Tool for Assembling Large Shotgun" http://www.jcvi.org/ cms/publications/ listing / abstract/browse/124/article/tigr-as.
[14] "TIGR Assembler: A New Tool for Assembling Large Shotgun"http://www.jcvi.org/cms/uploads/media/TIGR assembler.Pdf
[15] "Sequence Comparison Tools" http://www.springer.com / cda/content/document/cda_downloaddocument/9780387927374
[16] "HPC bioinformatics portal" http://biohpc-portal .hpcl.gwu.edu/

DOI: 03.LSCS.2013.5. 48